

NanoRTax, a real-time pipeline for taxonomic and diversity analysis of nanopore 16S rRNA amplicon sequencing data

Héctor Rodríguez-Perez

Research Unit Hospital Universitario Nuestra Señora de la Candelaria, Universidad de La Laguna

Laura Ciuffreda

Research Unit Hospital Universitario Nuestra Señora de la Candelaria <https://orcid.org/0000-0003-0034-8181>

Carlos Flores (✉ cflores@ull.edu.es)

Nuestra Señora de la Candelaria University Hospital: Hospital Universitario Nuestra Señora de la Candelaria <https://orcid.org/0000-0001-5352-069X>

Research Article

Keywords: Pipeline, sequencing, nanopore, real-time, microbiome

Posted Date: October 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-938802/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

The study of microbial communities and their applications have been leveraged by the advances in sequencing techniques and bioinformatics tools. The Oxford Nanopore Technologies long read sequencing by nanopores provides a portable and cost-efficient platform for sequencing assays opening the possibility of its application outside specialized environments and real-time analysis of data. To complement the existing efficient library preparation protocol with a streamlined analytic workflow, here we present NanoRTax, a nextflow pipeline for nanopore 16S rRNA amplicon data that features state-of-art taxonomic classification tools and real-time capability. The pipeline is paired with a web-based visual interface to enable user-friendly inspections of the experiment in progress.

Background

Since the adoption of next-generation sequencing (NGS) technologies, the continuous development of sequencing techniques and cost reductions have revolutionized the study of microbial communities [1]. The ever-growing availability of sequencing equipment in research laboratories and facilities has dramatically increased the number of metagenomic studies, databases, and bioinformatic tools [2,3]. Consequently, a wide range of applications has emerged in life and health sciences, such as the integration of sequencing approaches in clinical settings [4,5], where these methods can bolster the speed and sensitivity of the traditional microbial culturing and antibiotic susceptibility testing [6].

The introduction of third-generation sequencing technologies, such as that of Oxford Nanopore Technologies (ONT), has enabled the sequencing of long reads (>1 kbp) while providing a portable platform, which confers the ability to sequence samples even in a non-specialized environment [7,8]. In particular, ONT long reads can span complete transcripts or genes, and target sequences such as the full-length 16S rRNA gene for taxonomic classification of bacteria. Specifically, the increase in read length has led to a boost in taxonomic resolution and classification accuracy, making it possible to assign reads beyond the genus-level when performing pathogen identification or diversity analyses [9]. Besides this, ONT sequencing platforms also feature the unique possibility to access read data while being generated by an ongoing experiment. This characteristic along with the availability of rapid library preparation protocols has served to operate with turnaround times of less than 6 h, a dramatic decrease from the 48-72 h required for microbial culture approaches - emphasizing the potential of bringing a streamlined sequencing and real-time analysis to critical time response settings [10,11].

This challenge requires pairing rapid laboratory protocols with bioinformatic tools adapted for real-time workflows. Besides, taxonomic classifiers of long reads need to comprehensively evaluate the effect of tool and database selection in a real-time analysis scenario [12]. Here we present NanoRTax, a nextflow-based long-read pipeline for bacterial taxonomy classification and sample diversity analysis. The pipeline features the integration of state-of-art read classification methods, downstream analysis, and real-time capability to enable benchmarking of full-length 16S rRNA gene classification workflows while the

sequencing experiment is in progress. The pipeline is paired with an independent Dash web application which provides immediate access to taxonomic information, diversity statistics, and visualizations.

Results

To assess the usefulness of NanoRTax real-time analysis capability, we analysed the full-length 16S rRNA gene nanopore sequencing data from 31 tracheal aspirates from adult Intensive Care Unit (ICU) patients with non-pulmonary sepsis collected from a single medical-surgical ICU at sepsis diagnosis (within 8 h). We previously described that a reduction in genus-level bacterial lung diversity within 8 h of sepsis diagnosis is associated with ICU mortality, providing a potentially novel and early prognostic biomarker of non-pulmonary sepsis with better prognostic ability than other commonly used clinical scores (Guillen-Guio et al., 2020). We performed the re-analysis of this data using the NanoRTax complete workflow.

Table 1. NanoRTax software dependencies and versions.

Fastp	v0.20.1
Kraken2	v1.1.1
Centrifuge	v1.0.4 beta
blastn	v2.11.0
Taxonkit	v0.8.0

For each ICU sample, species-level diversity index metrics were calculated periodically from 5,000 to 100,000 reads to simulate different time periods of an ongoing sequencing experiment. Shannon diversity index calculated at species level for each time period was then compared between deceased and survivor patients based on Kraken2. BLAST classifications were used only for reference. The predictive value of the lung bacterial diversity index was assessed by fitting a linear model and calculating the Area Under the Curve (AUC) from Receiver Operating Characteristic (ROC) curves (**Figure 2**). We observed a reduction in the Shannon diversity index in deceased ICU patients compared to survivors as early as at < 2 h of the simulated sequencing experiment, which roughly corresponds to 5,000 reads (Wilcoxon test, $p=0.002$ and AUC=88.67% (Kraken2); 86.00% (BLAST)).

These results were essentially equivalent to those obtained in a simulated experiment collecting reads up to 24-48 h, roughly corresponding to 100,000 reads (Wilcoxon test, $p=0.005$ and AUC=88.67% (Kraken2); 86.00% (BLAST)).

Discussion

The strong association of a reduced lung bacterial diversity with a worse sepsis prognosis highlights the importance of host-microbial interactions and provides an early prognostic biomarker for sepsis. An early

sepsis response has been proven to be of paramount importance for positive patient outcomes and will remain relevant until novel drugs or interventions are demonstrated to be effective [13]. Applications in the context of diagnosis and mortality prediction have been explored recently, aiming to integrate not only sequencing information but also clinical data to enable better diagnosis, prognosis prediction, or entailment of treatment [14–16]. In this study, we simulated a realistic scenario of a real-time framework to predict ICU mortality in sepsis patients based on 16S rRNA gene sequencing experiments on lung samples paired with rapid analysis protocols, allowing us to draw the same conclusions as those from a complete 48 h sequencing dataset. Moreover, these results validated the previously observed lung dysbiosis association with mortality (Guillen-Guio et al., 2020) to the species level as a result of the higher taxonomic resolution achieved by sequencing of the full-length 16S rRNA genes. NanoRTax enables the immediate analysis of data while sequencing by implementing Kraken2 and Centrifuge rapid classifiers, which have been recently evaluated for long-read metagenomic profiling [17,18]. Additionally, the taxonomic assignment can be performed with BLAST to provide a framework to benchmark the tool in a real-time context or to evaluate Kraken2 and Centrifuge tools against a gold-standard BLAST classification. Our results also serve as a proof-of-concept of how real-time bioinformatic workflows could be useful to shorten the turnaround times in critical care settings and suggest its possible use for future research on early-response strategies for sepsis.

While NanoRTax was designed for full-length 16S rRNA gene taxonomic analysis of microbial samples, a focus on different amplification targets and the use of pipeline parametrization could take the application beyond bacterial profiling. Similar NanoRTax-based classification workflows can be proposed for the detection of fungal and viral infections [19,20], while non-taxonomic targeted amplicons can profile either specific antimicrobial resistance genes or entire gene panels such as the resistome [21]. Furthermore, continuous releases of the ONT sequencing chemistry and improvements in the basecalling algorithms are expected to positively impact taxonomic assignments using NanoRTax. Moreover, ONT hardware releases like the ONT Flongle sequencing adapter or the ONT Voltrax library preparation device can simplify rapid portable sequencing workflows by reducing the resources needed for the experimental protocols [22]. Enhanced portability and analytical speed directly benefit the in-situ assessment of microbial samples and confer relevance to the real-time bioinformatics tools described in this study. However, there are substantial practical challenges for routine taxonomic classification and metagenomics applications outside research practices [23]. Both analytical factors, such as sensitivity limitations due to genome size, pathogen load or ease of microorganism lysis [24], and sample factors, such as background contamination issues [25,26] can affect classification results in metagenomics studies. Bioinformatic analysis also turned out to be non-trivial since the completeness and accuracy of the ever-growing sequence databases and different approaches of taxonomic methods have demonstrated to have an important effect on results [3,27,28]. Thus, careful interpretation and constant benchmarking of analysis methods and databases will be key for taxonomic classification and metagenomic applications success [29].

Conclusions

We have developed NanoRTax, a bioinformatics pipeline to enable real-time taxonomic analysis of full-length 16S rRNA nanopore reads featuring multiple classification tools and immediate output visualization. We applied the NanoRTax workflow to the evaluation of 16S rRNA gene sequencing data of lung samples aimed to predict mortality in sepsis patients admitted to the ICU. Our results obtained from the analysis of very early sequencing data (within 2 h) support the benefits of implementing NGS-based assessments in this scenario. Despite this field is experiencing a fast development pace, we expect that routine clinical metagenomics will remain outside critical time-response scenarios until limitations are addressed. We anticipate that real-time bioinformatic analysis tools and implementations will be advancing concurrently with NGS development and applications.

Methods

NanoRTax is implemented in the Nextflow (Di Tommaso et al., 2017) workflow management system to enable efficient parallel execution and built-in integration of software dependencies using Docker containers and conda environments.

NanoRTax input consists of basecalled and demultiplexed FASTQ files following the structure of MinKNOW sequencing software output directories. The output path of an ongoing experiment can be specified for real-time analysis of newly generated FASTQ files. First, input sequences undergo a quality control step based on read length and quality filtering using Fastp (Chen et al., 2018). Then, taxonomic assignment is performed by one or more classifiers of choice between Kraken2 [32], Centrifuge [33], and BLAST [34]. Database and parameter selection for each tool can be specified via command line or in pre-loaded configuration files. The classification output is then processed to extract the full taxonomy for every classified read using Taxonkit [35]. This information is used in the next step to generate the NanoRTax final output that includes the sequence relative abundances, OTU tables, and diversity index calculations at different taxonomic levels. A report aggregation step is performed while new FASTQ sequence files are fed to the pipeline and further classified. This enables the synchronization of NanoRTax execution with the sequencing experiment and allows the inspection of partial results of the ongoing experiment.

For user-friendly visualization of the partial or complete outputs, the pipeline can be paired with an independent Python Dash web application, which serves as a dashboard to explore outputs in real-time. The interface integrates interactive summary tables and plots regarding quality control parameters, relative abundances with modifiable frequency cutoffs, and sample diversity index calculations over time. The general workflow of NanoRTax and software versions are detailed in **Figure 1** and **Table 1**.

Declarations

Availability of data and materials

NanoRTax Nextflow pipeline and Python Dash web application are freely available under MIT license on Github [36] (<https://github.com/genomicsITER/NanoRTax>). Nextflow engine requires Java 8 or later and dependencies are automatically built using Conda (4.10.1 or later) or Docker (1.6.2 or later).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Instituto de Salud Carlos III [PI14/00844, PI17/00610, and FI18/00230] and co-financed by the European Regional Development Funds, “A way of making Europe” from the European Union; Ministerio de Ciencia e Innovación [RTC-2017-6471-1, AEI/FEDER, UE]; Cabildo Insular de Tenerife [CGIEU0000219140]; Fundación Canaria Instituto de Investigación Sanitaria de Canarias [PIFUN48/18]; and by the agreement with Instituto Tecnológico y de Energías Renovables (ITER) to strengthen scientific and technological education, training, research, development and innovation in Genomics, Personalized Medicine and Biotechnology [OA17/008].

Author’s contributions

HRP coded NanoRTax pipeline and web application, performed bioinformatic analysis and drafted the manuscript. LC performed the microbiome samples library preparation, sequencing and was a major contributor in testing, data analysis and writing the manuscript. CF supervised the project concept and study design, contributed to data analysis and all drafts, and obtained funding. All authors read and approved the final manuscript.

References

1. Ciuffreda L, Rodríguez-Pérez H, Flores C. Nanopore sequencing and its application to the study of microbial communities. *Comput Struct Biotechnol J* [Internet]. The Authors; 2021;19:1497–511. Available from: <https://doi.org/10.1016/j.csbj.2021.02.020>
2. Forbes JD, Knox NC, Ronholm J, Pagotto F, Reimer A. Metagenomics: The next culture-independent game changer. *Front Microbiol.* 2017;8:1–21.
3. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform.* Oxford University Press; 2018;20:1125–39.
4. Chiu CY, Miller SA. Clinical metagenomics. *Nat Rev Genet* [Internet]. Springer US; 2019;20:341–55. Available from: <http://dx.doi.org/10.1038/s41576-019-0113-7>
5. Greninger AL. The challenge of diagnostic metagenomics. *Expert Rev Mol Diagn* [Internet]. Taylor & Francis; 2018;18:605–15. Available from: <https://doi.org/10.1080/14737159.2018.1487292>

6. Miao Q, Ma Y, Wang Q, Pan J, Zhang Y, Jin W, et al. Microbiological Diagnostic Performance of Metagenomic Next-generation Sequencing When Applied to Clinical Practice. *Clin Infect Dis*. 2018;67:S231–40.
7. Mitsuhashi S, Kryukov K, Nakagawa S, Takeuchi JS, Shiraishi Y, Asano K, et al. A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer. *Sci Rep*. 2017;7:5657.
8. Oliva M, Milicchio F, King K, Benson G, Boucher C, Prosperi M. Portable nanopore analytics: are we there yet? *Bioinformatics* [Internet]. 2020;36:4399–405. Available from: <https://doi.org/10.1093/bioinformatics/btaa237>
9. Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. *Gigascience*. BioMed Central Ltd.; 2016;5:4.
10. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol* [Internet]. *Genome Biology*; 2015;16:1–14. Available from: <http://dx.doi.org/10.1186/s13059-015-0677-2>
11. Parker J, Helmstetter AJ, Devey Di, Wilkinson T, Papadopoulos AST. Field-based species identification of closely-related plants using real-time nanopore sequencing. *Sci Rep*. 2017;7:1–8.
12. Escobar-Zepeda A, Godoy-Lozano EE, Raggi L, Segovia L, Merino E, Gutiérrez-Rios RM, et al. Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics. *Sci Rep*. 2018;8:1–13.
13. Kim H II, Park S. Sepsis: Early recognition and optimized treatment. *Tuberc Respir Dis (Seoul)*. 2019;82:6–14.
14. Islam MM, Nasrin T, Walther BA, Wu CC, Yang HC, Li YC. Prediction of sepsis patients using machine learning approach: A meta-analysis. *Comput Methods Programs Biomed* [Internet]. Elsevier B.V.; 2019;170:1–9. Available from: <https://doi.org/10.1016/j.cmpb.2018.12.027>
15. Delahanty RJ, Alvarez JA, Flynn LM, Sherwin RL, Jones SS. Development and Evaluation of a Machine Learning Model for the Early Identification of Patients at Risk for Sepsis. *Ann Emerg Med* [Internet]. American College of Emergency Physicians; 2019;73:334–44. Available from: <https://doi.org/10.1016/j.annemergmed.2018.11.036>
16. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. *Front Med*. 2021;8.
17. Nicholls SM, Quick JC, Tang S, Loman NJ. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience*. Oxford University Press; 2019;8.

18. Urban L, Holzer A, Baronas JJ, Hall MB, Braeuninger-Weimer P, Scherm MJ, et al. Freshwater monitoring by nanopore sequencing. *Elife*. 2021;10:1–27.
19. Jun K II, Oh B-L, Kim N, Shin JY, Moon J. Microbial diagnosis of endophthalmitis using nanopore amplicon sequencing. *Int J Med Microbiol* [Internet]. Elsevier GmbH; 2021;311:151505. Available from: <https://doi.org/10.1016/j.ijmm.2021.151505>
20. Wang M, Fu A, Hu B, Tong Y, Liu R, Liu Z, et al. Nanopore Targeted Sequencing for the Accurate and Comprehensive Detection of SARS-CoV-2 and Other Respiratory Viruses. *Small*. 2020;16.
21. Lanza VF, Baquero F, Martínez JL, Ramos-Ruíz R, González-Zorn B, Andremont A, et al. In-depth resistome analysis by targeted metagenomics. *Microbiome*. *Microbiome*; 2018;6:1–14.
22. Levy SE, Boone BE. Next-generation sequencing strategies. *Cold Spring Harb Perspect Med*. 2019;9:1–12.
23. Schlaberg R, Chiu CY, Miller S, Procop GW, Weinstock G. Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Arch Pathol Lab Med*. 2017;141:776–86.
24. Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, Van Doorn LJ, et al. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Front Microbiol*. 2019;10:1–9.
25. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. *PeerJ*. 2014;2014:1–7.
26. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol*. 2014;12:1–12.
27. Chen Q, Zobel J, Verspoor K. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: A descriptive study. *Database*. 2017;2017:1–16.
28. R. Marcelino V, Holmes EC, Sorrell TC. The use of taxon-specific reference databases compromises metagenomic classification. *BMC Genomics*. *BMC Genomics*; 2020;21:1–5.
29. Sun Z, Huang S, Zhang M, Zhu Q, Haiminen N, Carrieri AP, et al. Challenges in benchmarking metagenomic profilers. *Nat Methods* [Internet]. 2021;18:618–26. Available from: <https://doi.org/10.1038/s41592-021-01141-3>
30. DI Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 2017;35:316–9.
31. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34:i884–90.

32. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol. Genome Biology*; 2019;20:1–13.
33. Kim D, Song L, Breitwieser FP, Salzberg SL. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res. Cold Spring Harbor Laboratory Press*; 2016;26:1721–9.
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
35. Shen W, Ren H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J Genet Genomics [Internet]. Elsevier Limited and Science Press*; 2021; Available from: <https://doi.org/10.1016/j.jgg.2021.03.006>
36. Rodríguez-Pérez H, Ciuffreda L, Flores C. NanoRTax source code [Internet]. 2021. Available from: <https://github.com/genomicsITER/NanoRTax>

Figures

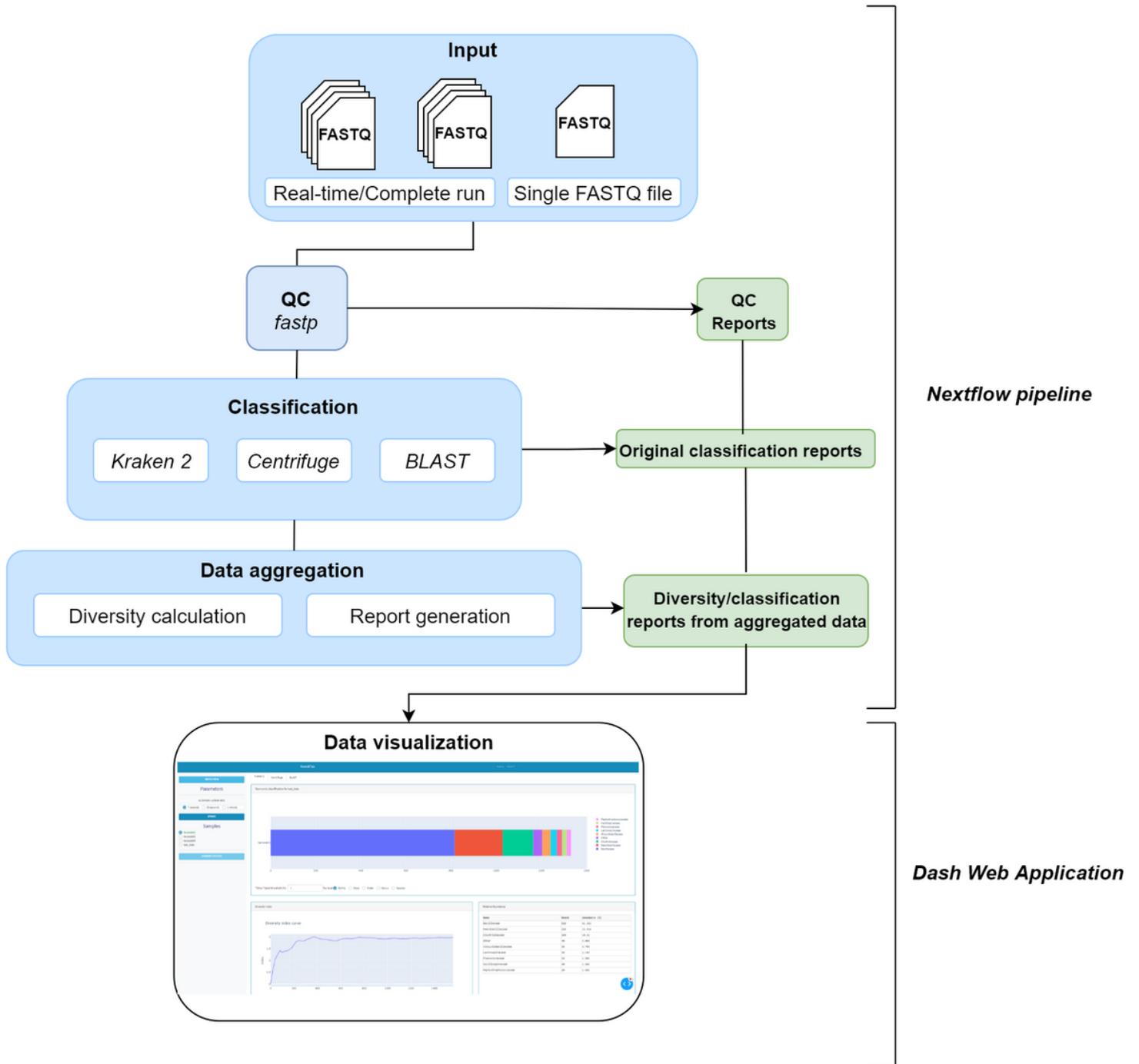


Figure 1

NanoRTax general overview.

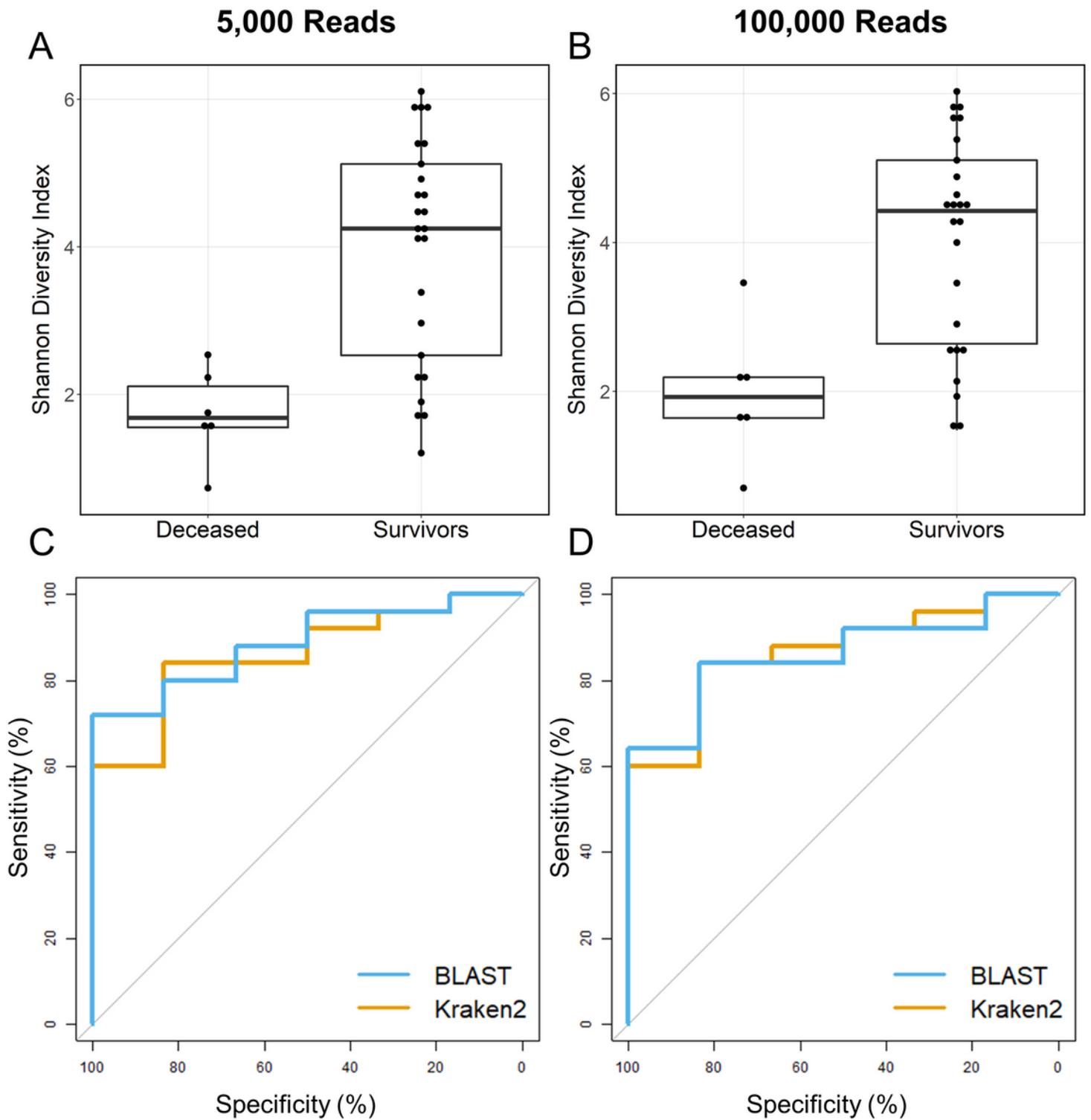


Figure 2

Boxplots of the Shannon diversity index per patient sample calculated with 5,000 reads (A) and 100,000 reads (B). Receiver Operating Characteristic (ROC) curves of mortality of ICU patients calculated with 5,000 reads (C) and 100,000 reads (D) for Kraken2 (orange) and BLAST (blue) classifiers at species level based on 8 h lung dysbiosis.